The intervention is possibly beneficial (and most unlikely harmful)

Profs Will Hopkins FACSM FECSS and Alan Batterham FACSM FRSS discuss magnitude-based inference and null-hypothesis significance testing.

An outcome with magnitude-based inference (MBI)

You've done the study, and the title of this article is the conclusion, using our approach to magnitude-based inference (MBI; Batterham & Hopkins, 2006; Hopkins *et al.,* 2009). Of course you will submit the study for publication, and you will expect your athletes, patients or clients to make use of the intervention, if it's not too expensive. OK, it's only *possibly* beneficial, but it's not going to harm them[a], so what have they got to lose? You or other researchers can do more studies of the intervention, and someone will eventually do a meta-analysis to reduce the uncertainty in the pooled mean effect.

Outcomes with null-hypothesis significance testing (NHST)

But wait. Your effect is not statistically significant ($p>0.05$). Now what? Do you want to change the conclusion? To what, there is no effect? Sorry, that's simply not true. If there is a good chance of benefit, it's absurd and unethical to claim there is no effect. And what about publishing the study? You'll have a hard time, if the reviewers are committed to null-hypothesis significance testing (NHST). The usual attitude of such reviewers is *significant = real = publishable,* and *non-significant = no effect = not publishable, unless the sample size is right*. That's the way NHST is meant to be used, and it works, sort-of. Amongst other problems with NHST, the right sample size is the one that gives statistical significance 80-90% of the time (the power of the study) for the smallest important beneficial effect, and for most studies in exercise and sport science it is impractically large. For example, with 80% power and 5% significance, a controlled trial of the effect of training with a new antioxidant on competitive endurance performance would need 350 competitive athletes in each group, and if you were looking at the effects of an injury-prevention programme on risk of injury, you would need *at least* 2,900 athletes in each group[b]. So if the reviewers are doing their job according to the precepts of NHST, your study and all other underpowered studies will not get into print with non-significant effects. Occasionally though, thanks to sampling variation, researchers doing underpowered studies fluke unrealistic big effects that turn out to be significant, and these studies *do* get published. Hence one of the main reasons we have publication bias: significant effects are inevitably bigger than non-significant effects.

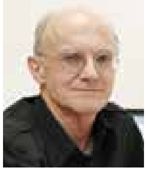
MBI vs NHST

Underpowered studies also occur in MBI, where the equivalent of *non-significant* is *unclear*, meaning too much uncertainty. But unclear effects are much less frequent than non-significant effects, so researchers using MBI get more of their studies published. What's more, publication bias with MBI is negligible. Altogether it's a no-brainer: MBI is superior to NHST. Unfortunately two traditional statisticians have recently tried to discredit MBI (Welsh & Knight, 2015). According to them, you can't say an effect is possibly beneficial unless you do a Bayesian analysis. MBI is actually a form of Bayesian analysis, but when we provided them with the published evidence (Batterham & Hopkins, 2015), they simply denied it. Their other main claim is that the Type-I error rate with MBI is unacceptably high in underpowered studies. A Type-I error occurs when a trivial true effect is declared substantial. In their analysis of non-clinical MBI, any overlap of a confidence interval with substantial values incurs a Type-I error for a null true effect, so they got rates of ~60%. In our analysis, a Type-I error occurs only when the confidence interval does not overlap trivial values, so the rate is at most 5%[c]. In simpler terms, if the true effect is trivial, you make a Type-I error in MBI only if you conclude that the effect is very unlikely to be trivial. Welsh's response (personal communication) is simply to deny our definition of a Type-I error. Stay tuned, and don't start putting p values back into your manuscripts just yet. ■

[a] That is, the mean effect in the population is possibly beneficial and most unlikely harmful. Individual responses to the intervention require a different analysis and should also be presented probabilistically. And if you're worried that possibly beneficial isn't likely enough for implementation, consider this: if you got $p=0.049$ for an intervention with the sample size that gives 80% or 90% power, it would actually be unlikely beneficial (20% or 10% chance of benefit).

[b] Sample sizes were estimated with the spreadsheet at Sportscience (Hopkins, 2006), assuming the smallest important change in performance is 0.3× the within-athlete variability in competitive performance, while the smallest important hazard ratio for injury is 1.11, with 50% incidence in the control group.

[c] For clinically important effects (those with the potential for benefit and harm), the Type-I error rate is higher, but it's generally less than that with NHST, and it's acceptable.



Prof Will Hopkins FACSM FECSS

Will is editor of Sportscience and has appointments with Victoria University, Melbourne, the Defence Institute, Oslo, and High Performance Sport NZ.

Prof Alan Batterham FACSM FRSS

Alan is at Teesside University and is Statistics Consultant/Advisor for the journals of The Physiological Society and BMJ Open.

References:

Batterham, A,M. & Hopkins, W.G. (2006). Making meaningful inferences about magnitudes. International Journal of Sports Physiology and Performance, 1, 50-57.

Batterham, A.M. & Hopkins, W.G. (2015). The case for magnitude-based inference. Medicine and Science in Sports and Exercise, 47, 885.

Hopkins, W.G. (2006). Estimating sample size for magnitude-based inferences. Sportscience, 10, 63-70.

Hopkins, W.G. *et al.* (2009). Progressive statistics for studies in sports medicine and exercise science. Medicine and Science in Sports and Exercise, 41, 3-12.

Welsh, A.H. & Knight, E.J. (2015). "Magnitude-based Inference": A statistical review. Medicine and Science in Sports and Exercise, 47, 874-884.